Implementation on Feature Subset Selection Using Symmetric Uncertainty Measure

Mrs.Jyoti Praful Gaidhani, Prof. S.B.Natikar

Abstract— Feature selection involves identifying a subset of most useful features that produces compatible results as the original entire set of features. In the process of Feature Selection the algorithm may be calculate from the efficiency and effectiveness points of view. The efficiency concerns the time required to get a subset of features and the effectiveness is related to the quality of subset of features. Using this criterion the clustering based feature selection algorithm is proposed and it uses computation of symmetric uncertainty measure between feature and target concept.
Feature Subset selection algorithm works in two steps. In first step, features are divided into clusters by using graph clustering methods. In second step the most representative feature which is strongly related to target class is selected from each cluster to form subset of features. Features in different clusters are independent; the clustering based strategy of algorithm has high probability of producing good subset of result

Index Terms— Feature Subset Selection, Filter Method, Feature Clustering, Graph Based clustering.

1 Introduction

Feature selection is a process that selects subset of original features. Feature subset selection is a preprocessing step to machine learning, is useful for reducing dimensionality, removing unrelated data, increasing learning accuracy & improving result comprehensibility. Due to increase in data dimensionality, there are many challenges for feature selection methods with respect to efficiency and effectiveness point of view. Feature subset selection method divided into four types Embedded, Wrapper, Filter and Hybrid.

The problem feature subset selection involves finding a good set of features under some objective functions. Common objective functions are prediction accuracy, Data dimensionality (structure size), and minimal use of input features. For finding the good subset of features it is necessary to identify relevant and redundant features on the basis of some criteria.

With respect to the filter feature selection methods, the cluster analysis and its application has been demonstrated to be more effective than traditional feature selection algorithms. In cluster analysis, graph-theoretic methods have been well studied and used in many applications.

By applying graph-theoretic clustering methods to features implement the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, a Fast clustering based feature selection algorithm (FAST) is proposed.

The FAST algorithm works in two steps. In the first, features are divided into clusters by using graph theoretic clustering methods. In the second step, the feature which is strongly related to target classes is selected from each cluster to form the final subset of features. Features are relatively independent in different clusters and this clustering based strategy of FAST has a high probability of producing a subset of useful and independent features.

2 Literature survey

2.1 Existing System

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other features. Of the many feature subset selection algorithms, some algorithm can effectively eliminate irrelevant features but fail to handle redundant features.

Devijver and Kittler [1] review feature selection methods for reducing the search space. But the heuristic causes a problem. Heuristic Search Algorithm performs poorly with feature interaction.

STAGGER [1] selects source features for constructing a new feature, judging from the feature weights based on their relevance to the concept. However, since relevance is determining one feature at a time, the method does not work for the domains where features interact with one another.

Relief [2] is a feature weight based algorithm and designed to pick all relevant features but does not help with redundant features.

The FOCUS [3] algorithm is able to detect the necessary and sufficient features in quasi-polynomial time. It is fast when target concept is very simple and data are noise free. But when data are noisy and complex then it is slow and selects many irrelevant features.

Consist [4] is also fast to compute and detect redundant as well as irrelevant features. It has been used with a variety of search strategies in feature selection and no modification is required. Consistency measure is monotonic, fast, able to remove redundant and irrelevant features and capable of handling some noise. Consistency measure can handle misclassifications.

CFS [5] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other.

FCBF [6] is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis.

Disadvantages

The selected feature has limited generality and the computational complexity is large.

Their computational complexity is minimum, but the accuracy of learning algorithms is not guaranteed.

3 PROPOSED SYSTEM

Different from these algorithms, proposed the fuzzy FAST employs the clustering based method for choosing the features.

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Quite different from these hierarchical clustering-based algorithms, proposed fuzzy FAST algorithm uses Graph theoretic clustering method to cluster features using fuzzy logic.

Graph Clustering is the best known graph theoretic divisive clustering algorithm which is based on construction of the minimal spanning tree (MST) of the data, and then deleting the MST edges with the largest lengths to generate clusters

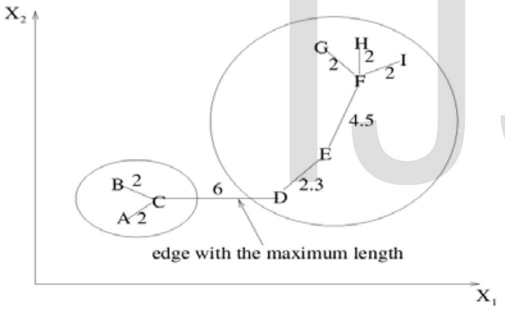


Fig.1 Clustering using Minimum Spanning Tree

Figure 1 depicts the MST obtained from two dimensional points. By cutting the link labeled CD with a length of 6 units (the edge with the maximum Euclidean length), two clusters (A, B, C) and (D, E, F, G, H, I) are obtained. The second cluster can be further divided into again two clusters by cutting the edge EF, which has a length of 4.5. The hierarchical approaches are also related to graph clustering. Single link clusters are subgraphs of the minimum spanning tree of the data which are also the connected components. Complete link clusters are maximal complete subgraphs, and are associated to the node colorability of graphs. The maximum complete subgraph was considered the strictest definition of a cluster.

Irrelevant features with redundant features, harshly affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, it develops a novel algorithm which can efficiently and effectively deals with both irrelevant and redundant features, achieve a good feature subset.

The following definitions used in this algorithm are as follows Suppose F to be the full set of features, Fi ∈ F be a feature, Si = F - {Fi} and $S'_i \subseteq$ Si. Let s'_i be a value-assignment of all features in S'_i , fi a value-assignment of feature Fi, and c a value-assignment of the target concept C.

Relevant Feature Fi is relevant to the target concept C if and only if there exists some s'_i, fi, and c, such that, for probability $p(S'_{i=} s'_i; Fi = fi) > 0$,

$p(C = c \mid S'_{i=} s'_i \mid Fi = fi) \neq p(C = c \mid S'_{i=} s'_i)$

Otherwise, feature Fi is an irrelevant feature.

Most of the information contained in redundant features is already present in other features. Redundant features do not contribute to getting better interpreting ability to the target concept.

The definitions of Markov blanket and redundant feature are introduced as follows, respectively.

Markov Blanket Given a feature Fi ∈ F, let Mi ⊂ F (Fi ∉ Mi), Mi is said to be a Markov blanket for Fi if and only if

p(F -Mi –{ Fi},C| Fi, Mi) =p(F -Mi – {Fi}, C | Mi)

Redundant Feature Let S be a set of features, a feature in S is redundant if and only if it has a Markov Blanket within S.

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes.

The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification by a number of researchers. Therefore, symmetric uncertainty use as the measure of correlation between either two features or a feature and the target concept.

The symmetric uncertainty(SU) [11] is given as

$$SU(X,Y) = \frac{2\times Gain(X|Y)}{H(X)+H(Y)}$$

Where,

1. H(X) is the entropy of a discrete random variable X.Suppose p(x) is the prior probabilities for all values of X, H(X) is defined by

$$H(X) = -\sum_{x\epsilon X} p(x)\log2\, p(x)$$

2. Gain(X|Y) is the amount by which the entropy of Y decreases. It reflects the additional information about Y provided by X and is called the information gain which is given by

Gain(X|Y) = H(X) – H(X|Y)
= H(X) – H(Y|X)

Where H(X|Y) is the conditional entropy which quantifies the remaining entropy (i.e., uncertainty) of a random variable X given that the value of another random variable Y is known. Suppose, p(x) is the prior probabilities for all values of X and p(x|y) is the posterior probabilities of X given the values of Y , H(X | Y) is defined by

$$H(X|Y) = -\sum_{y\epsilon Y} \mathrm{p(y)} \sum_{x\epsilon X} \mathrm{p(x|y)log2p(x|y)}$$

Information gain is a symmetrical measure. That means the amount of information gained about X after observing Y is equal to the amount of information gained about Y after observing X. This ensures that the order of two variables (e.g.,(X, Y) or (Y,X)) will not affect the value of the measure.

T-Relevance. The relevance between the feature Fi ϵ F and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by SU(Fi, C).
If SU(Fi, C) is greater than a predetermined threshold θ, we say that Fi is a strong T-Relevance feature.

F-Correlation The correlation between any pair of features Fi and Fj (Fi, Fj ϵ F $\wedge$ i $\neq$ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj).

F-Redundancy Let S ={F1, F2, . . . , Fi, . . . , $F_{k<|F|}$} be a cluster of features. if ∃ Fj ϵ S, SU(Fj,C)≥SU(Fi, C) ∧ SU(Fi,Fj) > SU(Fi, C) is always corrected for each Fi ∈ S (i≠j), then Fi are redundant features with respect to the given Fj (i.e., each Fi is a F-Redundancy).

R-Feature A feature Fi ∈ S ={F1, F2, . . . , Fk} (k< | F |) is a representative feature of the cluster S (i.e., Fi is a R-Feature) if and only if,
Fi =$\mathrm{argmax}_{Fi\in S}$SU(Fj,C)
It means the feature, which has the strongest T-Relevance, can act as a R-Feature for all the features in the cluster.
The proposed clustering based feature subset selection algorithm is given below.
The algorithm consists of three parts.
Part 1: It includes irrelevant feature removal.
Part 2: It includes construction of minimum spanning tree.
Part 3: It includes tree partition & representative feature selection using fuzzy logic.

Algorithm

Input: D(F1,F2,…..,Fm, C)-the given data set.
θ- the Target Relevance Threshold
Output: S-Selected feature subset

```
for  i=1 to m do
    T-Relevance= SU(Fi, C).
    if T-Relevance > θ then
        S=S ∪ { Fi }
G=NULL.//G is Complete Graph.
for each pair of features { $F'_i$, $F'_j$ }⊂S do
    F-Correlation = SU($F'_i$, $F'_j$)
    Add $F'_i$ and/or $F'_j$ to G with F-Correlation as the weight of the corresponding edge.
minSpanTree =Prim(G)
Forest = minSpanTree
for each edge $E_{ij}$∈ Forest do
    if SU($F'_i$,$F'_j$)< SU($F'_i$, C)∧ SU($F'_i$, $F'_j$)< SU($F'_j$, C) then
        Forest=Forest - $E_{ij}$
set the range for each tree R($R_{i........}R_m$)
Use Trapezoidal function on this range R.
S=Φ
for each tree $T_i$∈ Forest do
    $F_R^j$= $\mathrm{argmax}_{Fi\in Ti}$SU($F_k^i$,C)
    S=S∪{ $F_R^j$}
Return S.
```

Advantages:

Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other.

The efficiently and effectively deals with both irrelevant and redundant features for obtaining good feature subset.

4 SYSTEM DESIGN AND IMPLEMENTATION

System Architecture

The system architecture for clustering based feature selection is as follows

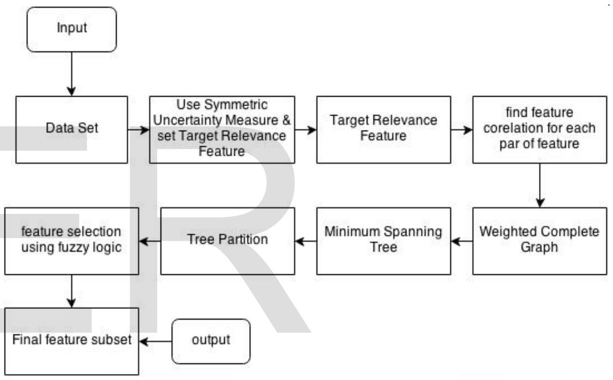

Input
Data Set
Use Symmetric Uncertainty Measure & set Target Relevance Feature
Target Relevance Feature
find feature corelation for each par of feature
Weighted Complete Graph
Minimum Spanning Tree
Tree Partition
feature selection using fuzzy logic
Final feature subset
output

Fig. 2. architecture of Fuzzy FAST

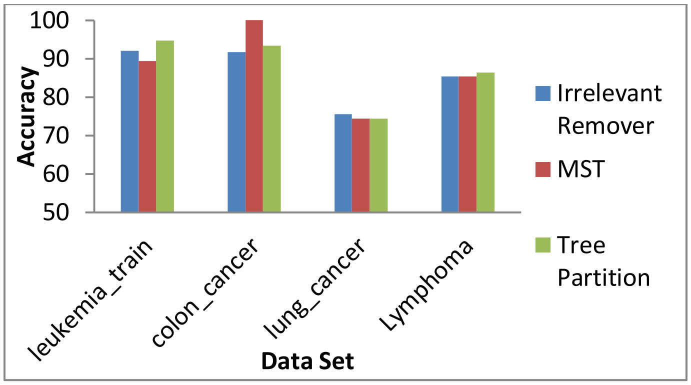
5 RESULT ANALYSIS

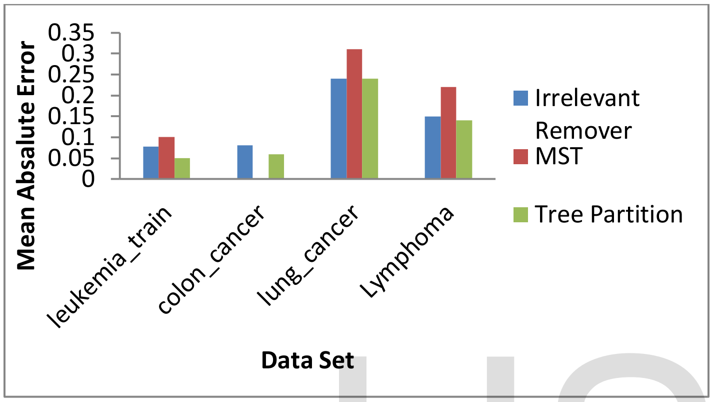
Data Set

Input to the system is following data set.

Luekemia,colon_cancer,Lung_cancer,Lymphoma.

The following graph shows result on accuracy and mean absolute error for the given data set.





6 CONCLUSION

Feature subset selection algorithm using concept of fuzzy logic is presented. This algorithm involves removing irrelevant features, building a minimum spanning tree from relative ones, and partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is severely reduced. The proposed algorithm obtain the better proportion of selected features, the better mean absolute error and the better classification accuracy.

REFERENCES

[1] Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[2] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[3] Fleuret, F. "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, vol. 5, pp. 1531-1555,2004.

[4] G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, pp. 121-129, 1994.

[5] K. Kira and L.A. Rendell, "The Feature Selection Problem:Traditional Methods and a New Algorithm," Proc. 10th Nat'l Conf.Artificial Intelligence, pp. 129-134, 1992.

[6] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf.Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[7] L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.

[8] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling,Numerical Recipes in C. Cambridge Univ. Press 1988

[9] H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence,vol. 69, nos. 1/2, pp. 279-305, 1994.